

How Good Are Sparse Cutting-Planes?*

Santanu S. Dey

Marco Molinaro

Qianyi Wang

May 9, 2014

Abstract

Sparse cutting-planes are often the ones used in mixed-integer programming (MIP) solvers, since they help in solving the linear programs encountered during branch-&-bound more efficiently. However, how well can we approximate the integer hull by just using sparse cutting-planes? In order to understand this question better, given a polyope P (e.g. the integer hull of a MIP), let P^k be its best approximation using cuts with at most k non-zero coefficients. We consider $d(P, P^k) = \max_{x \in P^k} (\min_{y \in P} \|x - y\|)$ as a measure of the quality of sparse cuts.

In our first result, we present general upper bounds on $d(P, P^k)$ which depend on the number of vertices in the polytope and exhibits three phases as k increases. Our bounds imply that if P has polynomially many vertices, using half sparsity already approximates it very well. Second, we present a lower bound on $d(P, P^k)$ for random polytopes that show that the upper bounds are quite tight. Third, we show that for a class of hard packing IPs, sparse cutting-planes do not approximate the integer hull well, that is $d(P, P^k)$ is large for such instances unless k is very close to n . Finally, we show that using sparse cutting-planes in extended formulations is at least as good as using them in the original polyhedron, and give an example where the former is actually much better.

1 Introduction

Most successful mixed integer linear programming (MILP) solvers are based on branch-&-bound and cutting-plane (cut) algorithms. Since MILPs belong to the class of NP-hard problems, one does not expect the size of branch-&-bound tree to be small (polynomial in size) for every instance. In the case where the branch-&-bound tree is not small, a large number of linear programs must be solved. It is well-known that dense cutting-planes are difficult for linear programming solvers to handle. Therefore, most commercial MILPs solvers consider sparsity of cuts as an important criterion for cutting-plane selection and use [6, 1, 10].

Surprisingly, very few studies have been conducted on the topic of sparse cutting-planes. Apart from cutting-plane techniques that are based on generation of cuts from single rows (which implicitly lead to sparse cuts if the underlying row is sparse), to the best of our knowledge only the paper [2] explicitly discusses methods to generate sparse cutting-planes.

The use of sparse cutting-planes may be viewed as a compromise between two competing objectives. As discussed above, on the one hand, the use of sparse cutting-planes aids in solving the linear programs encountered in the branch-&-bound tree faster. On the other hand, it is possible that ‘important’ facet-defining or valid inequalities for the convex hull of the feasible solutions are

*Santanu S. Dey and Qianyi Wang were partially supported by NSF grant CMMI-1149400.

dense and thus without adding these cuts, one may not be able to attain significant integrality gap closure. This may lead to a larger branch-&-bound tree and thus result in the solution time to increase.

It is challenging to simultaneously study both the competing objectives in relation to cutting-plane sparsity. Therefore, a first approach to understanding usage of sparse cutting-planes is the following: If we are able to separate and use valid inequalities with a given level of sparsity (as against completely dense cuts), how much does this cost in terms of loss in closure of integrality gap?

Considered more abstractly, the problem reduces to a purely geometric question: Given a polytope P (which represents the convex hull of feasible solutions of a MILP), how well is P approximated by the use of sparse valid inequalities. In this paper we will study polytopes contained in the $[0, 1]^n$ hypercube. This is without loss of generality since one can always translate and scale a polytope to be contained in the $[0, 1]^n$ hypercube.

1.1 Preliminaries

A cut $ax \leq b$ is called k -sparse if the vector a has at most k nonzero components. Given a set $P \subseteq \mathbb{R}^n$, define P^k as the best outer-approximation obtained from k -sparse cuts, that is, it is the intersection of all k -sparse cuts valid for P .

For integers k and n , let $[n] := \{1, \dots, n\}$ and let $\binom{[n]}{k}$ be the set of all subsets of $[n]$ of cardinality k . Given a k -subset of indices $I \subseteq [n]$, define $\mathbb{R}^{\bar{I}} = \{x \in \mathbb{R}^n : x_i = 0 \text{ for all } i \in I\}$. An equivalent and handy definition of P^k is the following: $P^k = \bigcap_{I \in \binom{[n]}{k}} (P + \mathbb{R}^{\bar{I}})$. Thus, if P is a polytope, then P^k is also a polytope.

1.2 Measure of Approximation

There are several natural measures to compare the quality of approximation provided by P^k in relation to P . For example, one may consider objective value ratio: maximum over all costs c of expression $\frac{z^{c,k}}{z^c}$, where $z^{c,k}$ is the value of maximizing c over P^k , and z^c is the same for P . We discard this ratio, since this ratio can become infinity and not provide any useful information. For example take $P = \text{conv}\{(0,0), (0,1), (1,1)\}$ and compare with P^1 wrt $c = (1, -1)$. Similarly, we may compare the volumes of P and P^k . However, this ratio is not useful if P is not full-dimensional and P^k is.

In order to have a useful measure that is well-defined for all polytopes contained in $[0, 1]^n$, we consider the following *distance measure*:

$$d(P, P^k) := \max_{x \in P^k} \left(\min_{y \in P} \|x - y\| \right),$$

where $\|\cdot\|$ is the ℓ_2 norm. It is easily verified that there is a vertex of P^k attaining the maximum above. Thus, alternatively the distance measure can be interpreted as the Euclidean distance between P and the farthest vertex of P^k from P .

Observation 1.1 ($d(P, P^k)$ is an upper bound on depth of cut). *Suppose $\alpha x \leq \beta$ is a valid inequality for P where $\|\alpha\| = 1$. Let the depth of this cut be the smallest $\gamma \geq 0$ such that $\alpha x \leq \beta + \gamma$ is valid for P^k . It is straightforward to verify that $\gamma \leq d(P, P^k)$. Therefore, the distance measure gives an upper bound on additive error when optimizing a (normalized) linear function over P and P^k .*

Observation 1.2 (Comparing $d(P, P^k)$ to \sqrt{n}). Notice that the largest distance between any two points in the $[0, 1]^n$ hypercube is at most \sqrt{n} . Therefore in the rest of the paper we will compare the value of $d(P, P^k)$ to \sqrt{n} .

1.3 Some Examples

In order to build some intuition we begin with some examples in this section. Let $P := \{x \in [0, 1]^n : ax \leq b\}$ where a is a non-negative vector. It is straightforward to verify that in this case, $P^k := \{x \in [0, 1]^n : a^I x \leq b \ \forall I \in \binom{[n]}{k}\}$, where $a_j^I := a_j$ if $j \in I$ and $a_j^I = 0$ otherwise.

Example 1: Consider the simplex $P = \{x \in [0, 1]^n : \sum_{i=1}^n x_i \leq 1\}$. Using the above observation, we have that $P^k = \text{conv}\{e^1, e^2, \dots, e^n, \frac{1}{k}e\}$, where e^j is the unit vector in the direction of the j^{th} coordinate and e is the all ones vector. Therefore the distance measure between P and P^k is $\sqrt{n}(\frac{1}{k} - \frac{1}{n}) \approx \frac{\sqrt{n}}{k}$, attained by the points $\frac{1}{n}e \in P$ and $\frac{1}{k}e \in P^k$. This is quite nice because with $k \approx \sqrt{n}$ (which is pretty reasonably sparse) we get a constant distance. Observe also that the *rate of change of the distance measure* follows a ‘single pattern’ - we call this a *single phase example*. See Figure 1(a) for $d(P, P^k)$ plotted against k (in blue) and $k \cdot d(P, P^k)$ plotted against k (in green).

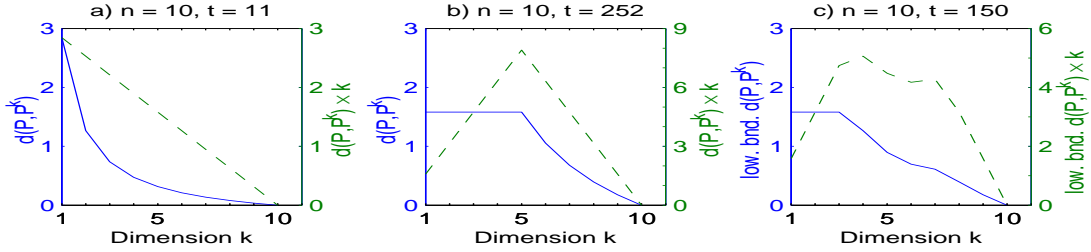


Figure 1: (a) Sparsity is good. (b) Sparsity is not so good. (c) Example with three phases.

Example 2: Consider the set $P = \{x \in [0, 1]^n : \sum_i x_i \leq \frac{n}{2}\}$. We have that $P^k := \{x \in [0, 1]^n : \sum_{i \in I} x_i \leq \frac{n}{2}, \ \forall I \in \binom{[n]}{k}\}$. Therefore, for all $k \in \{1, \dots, n/2\}$ we have $P^k = [0, 1]^n$ and hence $d(P, P^k) = \sqrt{n}/2$. Thus, we stay with distance $\Omega(\sqrt{n})$ (the worst possible for polytopes in $[0, 1]^n$) even with $\Theta(n)$ sparsity. Also observe that for $k > \frac{n}{2}$, we have $d(P, P^k) = \frac{n\sqrt{n}}{2k} - \frac{\sqrt{n}}{2}$. Clearly the rate of change of the distance measure has *two phases*, first phase of k between 1 and $\frac{n}{2}$ and the second phase of k between $\frac{n}{2}$ and n . See Figure 1(b) for the plot of $d(P, P^k)$ against k (in blue) and of $k \cdot d(P, P^k)$ against k (in green).

Example 3: We present an experimental example in dimension $n = 10$. The polytope P is now set as the convex hull of 150 binary points randomly selected from the hyperplane $\{x \in \mathbb{R}^{10} : \sum_{i=1}^{10} x_i = 5\}$. We experimentally computed lower bounds on $d(P, P^k)$ which are plotted in Figure 1(c) as the blue line (for details on this computation see Section B of the appendix). Notice that there are now *three phases*, which are more discernible in the plot between the lower bound on $k \cdot d(P, P^k)$ and k (in green).

The above examples serve to illustrate the fact that different polytopes, behave very differently when we try and approximate them using sparse inequalities.

2 Main Results

2.1 Upper Bounds

Surprisingly, it appears that the complicated behavior of $d(P, P^k)$ as k changes can be described to some extent in closed form. Our first result is a nontrivial upper bounds on $d(P, P^k)$ for general polytopes. This result is proven in Section 3.

Theorem 2.1 (Upper Bound on $d(P, P^k)$). *Let $n \geq 2$. Let $P \subseteq [0, 1]^n$ be the convex hull of points $\{p^1, \dots, p^t\}$. Then*

1. $d(P, P^k) \leq 4 \max \left\{ \frac{n^{1/4}}{\sqrt{k}} \sqrt{8 \max_{i \in [t]} \|p^i\| \sqrt{\log 4tn}}, \frac{8\sqrt{n}}{3k} \log 4tn \right\}$
2. $d(P, P^k) \leq 2\sqrt{n} \left(\frac{n}{k} - 1 \right).$

Since $\max_{i \in \{1, \dots, t\}} \|p^i\| \leq \sqrt{n}$ and the first upper bound yields nontrivial values only when $k \geq \frac{32}{3} \log 4tn$, a simpler (although weaker) expression for the first upper bound is $8\sqrt{2} \frac{\sqrt{n}}{\sqrt{k}} \sqrt{\log 4tn}$.¹ We make two observations based on Theorem 2.1.

Consider polytopes with ‘few’ vertices, say n^q vertices for some constant q . Suppose we decide to use cutting-planes with half sparsity (i.e. $k = \frac{n}{2}$), a reasonable assumption in practice. Then plugging in these values, it is easily verified that $d(P, P^k) \leq 16\sqrt{(q+1)\log n} \approx c\sqrt{\log n}$ for a constant c , which is a significantly small quantity in comparison to \sqrt{n} . In other words, *if the number of vertices is small, independent of the location of the vertices, using half sparsity cutting-planes allows us to approximate the integer hull very well*. We believe that as the number of vertices increase, the structure of the polytope becomes more important in determining $d(P, P^k)$ and Theorem 2.1 only captures the worst-case scenario. Overall, Theorem 2.1 presents a theoretical justification for the use of sparse cutting-planes in many cases.

Theorem 2.1 supports the existence of three phases in the behavior of $d(P, P^k)$ as k varies: **(Small k)** When $k \leq 128 \log 4tn$ the (simplified) upper bounds are larger than \sqrt{n} , indicating that ‘no progress’ is made in approximating the shape of P (this is seen Examples 2 and 3). **(Medium k)** When $128 \log 4tn \leq k \lesssim n - \sqrt{n \log 4tn}$ the first upper bound in Theorem 2.1 dominates. **(Large k)** When $k \gtrsim n - \sqrt{n \log 4tn}$ the upper bound $2\sqrt{n} \left(\frac{n}{k} - 1 \right)$ dominates. In particular, in this phase, $k \cdot d(P, P^k) \leq 2n^{3/2} - 2\sqrt{nk}$, i.e., the upper bound times k is a linear function of k . All the examples in Section 1 illustrate this behaviour.

2.2 Lower Bounds

How good is the quality of the upper bound presented in Theorem 2.1? Let us first consider the second upper bound in Theorem 2.1. Then observe that for the second example in Section 1, this upper bound is tight up to a constant factor for k between the values of $\frac{n}{2}$ and n .

We study lower bounds on $d(P, P^k)$ for random 0/1 polytopes in Section 4 that show that the first upper bound in Theorem 2.1 is also quite tight.

¹If $k \geq \frac{8 \log 4tn}{9}$, then $\frac{n^{1/4} \sqrt{8\sqrt{n} \sqrt{\log 4tn}}}{\sqrt{k}} \geq \frac{8\sqrt{n} \log 4tn}{3k}$

Theorem 2.2. *Let $k, t, n \in \mathbb{Z}_+$ satisfying $64 \leq k \leq n$ and $(0.5k^2 \log n + 2k + 1)^2 \leq t \leq e^n$. Let $\mathbf{X}^1, \mathbf{X}^2, \dots, \mathbf{X}^t$ be independent uniformly random points in $\{0, 1\}^n$ and let $\mathbf{P} = \text{conv}(\mathbf{X}^1, \mathbf{X}^2, \dots, \mathbf{X}^t)$. Then with probability at least $1/4$ we have that*

$$d(\mathbf{P}, \mathbf{P}^k) \geq \min \left\{ \frac{\sqrt{n}}{\sqrt{k}} \frac{\sqrt{\log t}}{110\sqrt{\log n}}, \frac{\sqrt{n}}{8} \right\} \left(\frac{1}{2} - \frac{1}{k^{3/2}} \right) - 3\sqrt{\log t}.$$

Let us compare this lower bound with the simpler expression $8\sqrt{2} \frac{\sqrt{n}}{\sqrt{k}} \sqrt{\log tn}$ for the first part of the upper bound of Theorem 2.1. We focus on the case where the minimum in the lower bound is achieved by the first term. Then comparing the leading term $\sqrt{\frac{n}{k}} \frac{\sqrt{\log t}}{2 \cdot 110 \sqrt{\log n}}$ in the lower bound with the upper bound, we see that these quantities match up to a factor of $O\left(\frac{\sqrt{\log(tn)}\sqrt{\log n}}{\sqrt{\log t}}\right)$, showing that for many 0/1 polytopes the first upper bound of Theorem 2.1 is quite tight. We also remark that in order to simplify the exposition we did not try to optimize constants and lower order terms in our bounds.

The main technical tool for proving this lower bound is a new anticoncentration result for linear combinations $a\mathbf{X}$, where the \mathbf{X}_i 's are independent Bernoulli random variables (Lemma 4.2). The main difference from standard anticoncentration results is that the latter focus on variation around the standard deviation; in this case, standard tools such as the Berry-Esseen Theorem or the Paley-Zygmund Inequality[4] can be used to obtain constant-probability anticoncentration. However, we need to control the behavior of $a\mathbf{X}$ much further away from its standard deviation, where we cannot hope to get constant-probability anticoncentration.

2.3 Hard Packing Integer Programs

We also study well-known, randomly generated, hard packing integer program instances (see for instance [7]). Given parameters $n, m, M \in \mathbb{N}$, the convex hull of the packing IP is given by $\mathbf{P} = \text{conv}(\{x \in \{0, 1\}^n : \mathbf{A}^j x \leq \frac{\sum_i \mathbf{A}_i^j}{2}, \forall j \in [m]\})$, where the \mathbf{A}_i^j 's are chosen independently and uniformly in the set $\{0, 1, \dots, M\}$. Let (n, m, M) -PIP denote the distribution over the generated \mathbf{P} 's.

The following result shows the limitation of sparse cuts for these instances.

Theorem 2.3. *Consider $n, m, M \in \mathbb{N}$ such that $n \geq 50$ and $8 \log 8n \leq m \leq n$. Let \mathbf{P} be sampled from the distribution (n, m, M) -PIP. Then with probability at least $1/2$, $d(\mathbf{P}, \mathbf{P}^k) \geq \frac{\sqrt{n}}{2} \left(\frac{2}{\max\{\alpha, 1\}} (1 - \epsilon)^2 - (1 + \epsilon') \right)$, where $c = k/n$ and*

$$\frac{1}{\alpha} = \frac{M}{2(M+1)} \left[\frac{n - 2\sqrt{n \log 8m}}{c((2-c)n + 1) + 2\sqrt{10cnm}} \right], \quad \epsilon = \frac{24\sqrt{\log 4n^2 m}}{\sqrt{n}},$$

$$\epsilon' = \frac{3\sqrt{\log 8n}}{\sqrt{m} - 2\sqrt{\log 8n}}.$$

Notice that when m is sufficiently large, and n reasonably larger than m , we have ϵ and ϵ' approximately 0, and the above bound reduces to approximately $\frac{\sqrt{n}}{2} \left(\left(\frac{M}{M+1} \right) \left(\frac{n}{k(2-n/k)} \right) - 1 \right) \approx \frac{\sqrt{n}}{2} \left(\frac{n}{k(2-n/k)} - 1 \right)$, which is within a constant factor of the upper bound from Theorem 2.1. The poor behavior of sparse cuts gives an indication for the hardness of these instances and suggests that denser cuts should be explored in this case.

One interesting feature of this result is that it works directly with the IP formulation, not relying on an explicit linear description of the convex hull.

2.4 Sparse Cutting-Planes and Extended Formulations

Let $\text{proj}_x : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ denote the projection operator onto the first n coordinates. We say that a set $Q \subseteq \mathbb{R}^n \times \mathbb{R}^m$ is an *extended formulation* of $P \subseteq \mathbb{R}^n$ if $P = \text{proj}_x(Q)$.

As our final result we remark that using sparse cutting-planes in extended formulations is at least as good as using them in the original polyhedron, and sometime much better. These results are proved in Section 6.

Proposition 2.4. *Consider a polyhedron $P \subseteq \mathbb{R}^n$ and an extended formulation $Q \subseteq \mathbb{R}^n \times \mathbb{R}^m$ for it. Then $\text{proj}_x(Q^k) \subseteq (\text{proj}_x(Q))^k = P^k$.*

Proposition 2.5. *Consider $n \in \mathbb{N}$ and assume it is a power of 2. Then there is a polytope $P \subseteq \mathbb{R}^n$ such that:*

1. $d(P, P^k) = \sqrt{n/2}$ for all $k \leq n/2$.
2. There is an extended formulation $Q \subseteq \mathbb{R}^n \times \mathbb{R}^{2n-1}$ of P such that $\text{proj}_x(Q^3) = P$.

3 Upper Bound

In this section we prove Theorem 2.1. In fact we prove the same bound for polytopes in $[-1, 1]^n$, which is a slightly stronger result. The following well-known property is crucial for the constructions used in both parts of the theorem.

Observation 3.1 (Section 2.5.1 of [3]). *Consider a compact convex set $S \subseteq \mathbb{R}^n$. Let \bar{x} be a point outside S and let \bar{y} be the closest point to \bar{x} in S . Then setting $a = \bar{x} - \bar{y}$, the inequality $ax \leq a\bar{y}$ is valid for S and cuts \bar{x} off.*

3.1 Proof of First Part of Theorem 2.1

Consider a polytope $P = \text{conv}\{p^1, p^2, \dots, p^t\}$ in $[-1, 1]^n$. Define

$$\lambda^* = \max \left\{ \frac{n^{1/4}}{\sqrt{k}} \sqrt{8 \max_i \|p^i\|} \sqrt{\log 4tn}, \frac{8\sqrt{n}}{3k} \log 4tn \right\}.$$

In order to show that $d(P, P^k)$ is at most $4\lambda^*$ we show that every point at distance more than $4\lambda^*$ from P is cut off by a valid inequality for P^k . Assume until the end of this section that $4\lambda^*$ is at most \sqrt{n} , otherwise the result is trivial; in particular, this implies that the second term in the definition of λ^* is at most $\sqrt{n}/4$ and hence $k \geq 8 \log 4tn$ (In fact, $k \geq \frac{32}{3} \log 4tn$, but $k \geq 8 \log 4tn$ suffices for the rest of the proof).

So let $u \in \mathbb{R}^n$ be a point at distance more than $4\lambda^*$ from P . Let $v \in P$ be the closest point in P to u . We can write $u = v + \lambda d$ for some vector d with $\|d\|_2 = 1$ and $\lambda > 4\lambda^*$. From Observation 3.1, inequality $dx \leq dv$ is valid for P , so in particular $dp^i \leq dv$ for all $i \in [t]$; in addition, it that this inequality cuts off u : $du = dv + \lambda > dv$. The idea is to use this extra slack factor λ in the previous equation to show we can ‘sparsify’ the inequality $dx \leq dv$ while maintaining separation of P and u . It then suffices to prove the following lemma.

Lemma 3.2. *There is a k -sparse vector $\tilde{d} \in \mathbb{R}^n$ such that*

1. $\tilde{d}p^i \leq \tilde{d}v + \frac{\lambda}{2}$, for all $i \in [t]$
2. $\tilde{d}u > \tilde{d}v + \frac{\lambda}{2}$.

To prove the lemma we construct a random vector $\tilde{\mathbf{D}} \in \mathbb{R}^n$ which, with non-zero probability, is k -sparse and satisfies the two other requirements of the lemma. Let $\alpha = \frac{k}{2\sqrt{n}}$. Define $\tilde{\mathbf{D}}$ as the random vector with independent coordinates, where $\tilde{\mathbf{D}}_i$ is defined as follows: if $\alpha|d_i| \geq 1$, then $\tilde{\mathbf{D}}_i = d_i$ with probability 1; if $\alpha|d_i| < 1$, then $\tilde{\mathbf{D}}_i$ takes value $\text{sign}(d_i)/\alpha$ with probability $\alpha|d_i|$ and takes value 0 with probability $1 - \alpha|d_i|$. (For convenience we define $\text{sign}(0) = 1$.)

The next proposition follows directly from the definition of $\tilde{\mathbf{D}}$.

Proposition 3.3. *For every vector $a \in \mathbb{R}^n$ the following hold:*

1. $\mathbb{E}[\tilde{\mathbf{D}}a] = da$
2. $\text{Var}(\tilde{\mathbf{D}}a) \leq \frac{1}{\alpha} \sum_{i \in [n]} a_i^2 |d_i|$
3. $|\tilde{\mathbf{D}}_i a_i - \mathbb{E}[\tilde{\mathbf{D}}_i a_i]| \leq \frac{|a_i|}{\alpha}$.

Claim 3.4. *With probability at least $1 - 1/4n$, $\tilde{\mathbf{D}}$ is k -sparse.*

Proof. Construct the vector $a \in \mathbb{R}^n$ as follows: if $\alpha|d_i| \geq 1$ then $a_i = 1/d_i$, and if $\alpha|d_i| < 1$ then $a_i = \alpha/\text{sign}(d_i)$. Notice that $\tilde{\mathbf{D}}a$ equals the number of non-zero coordinates of $\tilde{\mathbf{D}}$ and $\mathbb{E}[\tilde{\mathbf{D}}a] \leq \alpha\|d\|_1 \leq k/2$. Here the first inequality follows from the fact that $\mathbb{E}(\tilde{\mathbf{D}}_i a_i) \leq \alpha|d_i|$ for all i and the second inequality follows from the definition of α and the fact that $\|d\|_2 = 1$. Also, from Proposition 3.3 we have

$$\text{Var}(\tilde{\mathbf{D}}a) \leq \frac{1}{\alpha} \sum_{i \in [n]} a_i^2 |d_i| \leq \alpha\|d\|_1 \leq \frac{k}{2}.$$

Then using Bernstein's inequality (Section A of the appendix) we obtain

$$\Pr(\tilde{\mathbf{D}}a > k) \leq \exp\left(-\min\left\{\frac{k^2}{8k}, \frac{3k}{8}\right\}\right) \leq \frac{1}{4n},$$

where the last inequality uses our assumption that $k \geq 8 \log 4tn$. \square

We now show that property 1 required by Lemma 3.2 holds for $\tilde{\mathbf{D}}$ with high probability. Since $d(p^i - v) \leq 0$ for all $i \in [t]$, the following claim shows that property 1 holds with probability at least $1 - \frac{1}{4}$.

Claim 3.5. $\Pr(\max_{i \in [t]} [\tilde{\mathbf{D}}(p^i - v) - d(p^i - v)] > 2\lambda^*) \leq 1/4n$.

Proof. Define the centered random variable $\mathbf{Z} = \tilde{\mathbf{D}} - d$. To make the analysis cleaner, notice that $\max_{i \in [t]} \mathbf{Z}(p^i - v) \leq 2 \max_{i \in [t]} |\mathbf{Z}p^i|$; this is because $\max_{i \in [t]} \mathbf{Z}(p^i - v) \leq \max_{i \in [t]} |\mathbf{Z}p^i| + |\mathbf{Z}v|$, and because for all $a \in \mathbb{R}^n$ we have $|av| \leq \max_{p \in P} |ap| = \max_{i \in [t]} |ap^i|$ (since $v \in P$).

Therefore our goal is to upper bound the probability that the process $\max_{i \in [t]} |\mathbf{Z}p^i|$ is larger than λ^* . Fix $i \in [t]$. By Bernstein's inequality,

$$\Pr(|\mathbf{Z}p^i| > \lambda^*) \leq \exp\left(-\min\left\{\frac{(\lambda^*)^2}{4\text{Var}(|\mathbf{Z}p^i|)}, \frac{3\lambda^*}{4M}\right\}\right), \quad (1)$$

where M is an upper bound on $\max_j |\mathbf{Z}_j p_j^i|$.

To bound the terms in the right-hand side, from Proposition 3.3 we have

$$\text{Var}(\mathbf{Z}p^i) = \text{Var}(\tilde{\mathbf{D}}p^i) \leq \frac{1}{\alpha} \sum_j (p_j^i)^2 |d_j| \leq \frac{1}{\alpha} \sum_j p_j^i |d_j| \leq \frac{1}{\alpha} \|p^i\| \|d\| = \frac{1}{\alpha} \|p^i\|,$$

where the second inequality follows from the fact $p^i \in [0, 1]^n$, and the third inequality follows from the Cauchy-Schwarz inequality. Moreover, it is not difficult to see that for every random variable \mathbf{W} , $\text{Var}(|\mathbf{W}|) \leq \text{Var}(\mathbf{W})$. Using the first term in the definition of λ^* , we then have

$$\frac{(\lambda^*)^2}{\text{Var}(|\mathbf{Z}p^i|)} \geq 4 \log 4tn.$$

In addition, for every coordinate j we have $|\mathbf{Z}_j p_j^i| = |\tilde{\mathbf{D}}_j p_j^i - \mathbb{E}[\tilde{\mathbf{D}}_j p_j^i]| \leq 1/\alpha$, where the inequality follows from Proposition 3.3. Then we can set $M = 1/\alpha$ and using the second term in the definition of λ^* we get $\frac{\lambda^*}{M} \geq \frac{4}{3} \log 4tn$. Therefore, replacing these bounds in inequality (1) gives $\Pr(|\mathbf{Z}p^i| \geq \lambda^*) \leq \frac{1}{4tn}$.

Taking a union bound over all $i \in [t]$ gives that $\Pr(\max_{i \in [t]} |\mathbf{Z}p^i| \geq \lambda^*) \leq 1/4n$. This concludes the proof of the claim. \square

Claim 3.6. $\Pr(\tilde{\mathbf{D}}(u - v) \leq \lambda/2) \leq 1 - 1/(2n - 1)$.

Proof. Recall $u - v = \lambda d$, hence it is equivalent to bound $\Pr(\tilde{\mathbf{D}}d \leq 1/2)$. First, $\mathbb{E}[\tilde{\mathbf{D}}d] = dd = 1$. Also, from Proposition 3.3 we have $\tilde{\mathbf{D}}d \leq |\tilde{\mathbf{D}}d - dd| + |dd| \leq \frac{1}{\alpha} \sum_i |d_i| + 1 \leq \frac{2n}{k} + 1 \leq n$, where the last inequality uses the assumption $k \geq 8 \log 4tn$. Then employing Markov's inequality to the non-negative random variable $n - \tilde{\mathbf{D}}d$, we get $\Pr(\tilde{\mathbf{D}}d \leq 1/2) \leq 1 - \frac{1}{2n-1}$. This concludes the proof. \square

Proof of Lemma 3.2. Employ the previous three claims and union bound to find a realization of $\tilde{\mathbf{D}}$ that is k -sparse and satisfies requirements 1 and 2 of the lemma.

This concludes the proof of the first part of Theorem 2.1.

Observation 3.7. Notice that in the above proof λ^* is set by Claim 3.5, and need to be essentially $\mathbb{E}[\max_{i \in [t]} (\tilde{\mathbf{D}} - d)p^i]$. There is a vast literature on bounds on the supremum of stochastic processes (see for instance [8]), and improved bounds for structured P 's are possible (for instance, via the generic chaining method).

3.2 Proof of Second Part of Theorem 2.1

The main tool for proving this upper bound is the following lemma, which shows that when P is 'simple', and we have a stronger control over the distance of a point \bar{x} to P , then there is a k -sparse inequality that cuts \bar{x} off.

Lemma 3.8. Consider a halfspace $H = \{x \in \mathbb{R}^n : ax \leq b\}$ and let $P = H \cap [-1, 1]^n$. Let $\bar{x} \in [-1, 1]^n$ be such that $d(\bar{x}, H) > 2\sqrt{n}(\frac{n}{k} - 1)$. Then $\bar{x} \notin P^k$.

Proof. Assume without loss of generality that $\|a\|_2 = 1$. Let \bar{y} be the point in H closest to \bar{x} , and notice that $\bar{x} = \bar{y} + \lambda a$ where $\lambda > \sqrt{n}(\frac{n}{k} - 1)$.

For any set $I \in \binom{[n]}{k}$, the inequality $\sum_{i \in I} a_i x_i \leq b + \sum_{i \notin I: a_i \geq 0} a_i - \sum_{i \notin I: a_i < 0} a_i$ is valid for P ; since it is k -sparse, it is also valid for P^k . Averaging out these inequalities over all $I \in \binom{[n]}{k}$, we get that the following is valid for P^k :

$$\frac{k}{n} ax \leq b + \left(1 - \frac{k}{n}\right) \left(\sum_{i: a_i \geq 0} a_i - \sum_{i: a_i < 0} a_i \right) \equiv ax \leq b + \left(\frac{n}{k} - 1\right) (b + \|a\|_1).$$

We claim that \bar{x} violates this inequality. First notice that $a\bar{x} = a\bar{y} + \lambda = b + \lambda > b + 2\sqrt{n}(\frac{n}{k} - 1)$, hence it suffices to show $b + \|a\|_1 \leq 2\sqrt{n}$. Our assumption on \bar{x} implies that $P \neq [-1, 1]^n$, and hence $b < \max_{x \in [-1, 1]^n} ax = \|a\|_1$; this gives $b + \|a\|_1 \leq 2\|a\|_1 \leq 2\sqrt{n}\|a\|_2 = 2\sqrt{n}$, thus concluding the proof. \square

To prove the second part of Theorem 2.1 consider a point \bar{x} of distance greater than $2\sqrt{n}(\frac{n}{k} - 1)$ from P ; we show $\bar{x} \notin P^k$. Let \bar{y} be the closest point to \bar{x} in P . Let $a = \bar{x} - \bar{y}$. From Observation 3.1 we have that $ax \leq a\bar{y}$ is valid for P . Define $H' = \{x \in \mathbb{R}^n : ax \leq a\bar{y}\}$ and $P' = H' \cap [-1, 1]^n$. Notice that $d(\bar{x}, H') = d(\bar{x}, \bar{y}) > 2\sqrt{n}(\frac{n}{k} - 1)$. Then Lemma 3.8 guarantees that \bar{x} does not belong to P'^k . But $P \subseteq P'$, so by monotonicity of the k -sparse closure we have $P^k \subseteq P'^k$; this shows that $\bar{x} \notin P^k$, thus concluding the proof.

4 Lower Bound

In this section we prove Theorem 2.2. The proof is based on the ‘bad’ polytope of Example 2. For a random polytope \mathbf{Q} in \mathbb{R}^n , it is useful to think of each of its (random) faces from the perspective of supporting hyperplanes: for a fixed direction $d \in \mathbb{R}^n$, we have the valid inequality $dx \leq \mathbf{d}_0$, where $\mathbf{d}_0 = \max_{q \in \mathbf{Q}} dq$.

The idea of the proof is then to proceed in two steps. First, for a uniformly random 0/1 polytope \mathbf{P} , we show that with good probability the faces $dx \leq \mathbf{d}_0$ for \mathbf{P}^k have \mathbf{d}_0 being large, namely $\mathbf{d}_0 \gtrsim \left(\frac{1}{2} + \frac{\sqrt{\log t}}{\sqrt{k}}\right) \sum_i d_i$, forced by some point $p \in \mathbf{P}$ with large dp ; therefore, with good probability the point $\bar{p} \approx \left(\frac{1}{2} + \frac{\sqrt{\log t}}{\sqrt{k}}\right)e$ belongs to \mathbf{P}^k . In the second step, we show that with good probability the distance from \bar{p} to \mathbf{P} is at least $\approx \sqrt{\frac{n}{k}}\sqrt{\log t}$, by showing that the inequality $\sum_i x_i \lesssim \frac{n}{2} + \sqrt{n}$ is valid for \mathbf{P} .

We now proceed with the proof. Assume the conditions on k, n, t as stated in Theorem 2.2 hold. Consider the random set \mathcal{X} defined as $\{\mathbf{X}^1, \mathbf{X}^2, \dots, \mathbf{X}^t\}$ where the \mathbf{X}^i 's are independent uniform random points in $\{0, 1\}^n$, and define the random 0/1 polytope $\mathbf{P} = \text{conv}(\mathcal{X})$. To formalize the preceding discussion, we need the following definition.

Definition 4.1. We say that a (deterministic) 0/1 polytope in \mathbb{R}^n is α -tough if for every facet $dx \leq d_0$ of its k -sparse closure we have $d_0 \geq \frac{\sum_i d_i}{2} + \frac{\alpha}{2\sqrt{k}}(1 - \frac{1}{k^2})\|d\|_1 - \|d\|_\infty/2k^2$, for every $k \in \{2, \dots, n\}$.

The main element of the lower bound is the following anticoncentration result; in our setting, the idea is that for every (k -sparse) direction $d \in \mathbb{R}^n$, with good probability we will have a point p in \mathbf{P}^k (in fact in \mathbf{P}) with large dp .

Lemma 4.2. Let $\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_n$ be independent random variables with \mathbf{Z}_i taking value 0 with probability 1/2 and value 1 with probability 1/2 for every $i \in [n]$. Then for every $a \in [-1, 1]^n$ and $\alpha \in [0, \frac{\sqrt{n}}{8}]$,

$$\Pr\left(a\mathbf{Z} \geq \mathbb{E}[a\mathbf{Z}] + \frac{\alpha}{2\sqrt{n}} \left(1 - \frac{1}{n^2}\right) \|a\|_1 - \frac{1}{2n^2}\right) \geq \left(e^{-50\alpha^2} - e^{-100\alpha^2}\right)^{60 \log n}.$$

The proof of this lemma is reasonably simple and proceeds by grouping the random variables with similar a_i 's and then applies known anticoncentration to each of these groups; this proof is presented in Section C of the appendix.

In order to effectively apply this anticoncentration to *all* valid inequalities/directions of \mathbf{P}^k , we need some additional control. Define $\mathcal{D} \subseteq \mathbb{Z}^n$ as the set of all integral vectors $\ell \in \mathbb{R}^n$ that are k -sparse and satisfy $\|\ell\|_\infty \leq (k)^{k/2}$.

Lemma 4.3. Let $Q \subseteq \mathbb{R}^n$ be a 0/1 polytope. Then for every $k \in [n]$, there is a subset $\mathcal{D}' \subseteq \mathcal{D}$ such that $Q^k = \{x : dx \leq \max_{y \in Q^k} dy, d \in \mathcal{D}'\}$.

This lemma follows directly from applying Corollary 26 in [11] to each term $Q + \mathbb{R}^{\bar{I}}$ in the definition of Q^k from Section 1.1.

Employing this lemma to each scenario, we get that all the directions of facets of \mathbf{P}^k come from the set \mathcal{D} . This allows us to analyze the probability that \mathbf{P} is α -tough.

Lemma 4.4. Assume the conditions on k, n, t as stated in Theorem 2.2 hold. If $1 \leq \alpha^2 \leq \min\left\{\frac{\log t}{12000 \log n}, \frac{k}{64}\right\}$, then \mathbf{P} is α -tough with probability at least 1/2.

Proof. Let \mathcal{E} be the event that for all $d \in \mathcal{D}$ we have $\max_{i \in [t]} d\mathbf{X}^i \geq \frac{1}{2} \sum_j d_j + \frac{\alpha}{2\sqrt{k}}(1 - \frac{1}{k^2})\|d\|_1 - \|d\|_\infty/2k^2$. Because of Lemma 4.3, whenever \mathcal{E} holds we have that \mathbf{P} is α -tough and thus it suffices to show $\Pr(\mathcal{E}) \geq 1/2$.

Fix $d \in \mathcal{D}$. Since d is k -sparse, and $\alpha \leq \frac{\sqrt{k}}{8}$, we can apply Lemma 4.2 to $d/\|d\|_\infty$ restricted to the coordinates in its support to obtain that

$$\begin{aligned} \Pr\left(d\mathbf{X}^i \geq \frac{\sum_i d_i}{2} + \frac{\alpha}{2\sqrt{k}} \left(1 - \frac{1}{k^2}\right) \|d\|_1 - \frac{\|d\|_\infty}{2k^2}\right) &\geq \left(e^{-50\alpha^2} - e^{-100\alpha^2}\right)^{60 \log n} \\ &\geq e^{-100\alpha^2 \cdot 60 \log n} \geq \frac{1}{t^{1/2}}, \end{aligned}$$

where the second inequality follows from the lower bound on α^2 (in fact $\alpha^2 \geq \frac{\log 2}{50}$ is sufficient) and the last inequality follows from our upper bound on α^2 . By independence of the \mathbf{X}^i 's,

$$\Pr\left(\max_{i \in [t]} d\mathbf{X}^i < \frac{\sum_i d_i}{2} + \frac{\alpha}{2\sqrt{k}} \left(1 - \frac{1}{k^2}\right) \|d\|_1 - \frac{\|d\|_\infty}{2k^2}\right) \leq \left(1 - \frac{1}{t^{1/2}}\right)^t \leq e^{-t^{1/2}},$$

where the second inequality follows from the fact that $(1 - x) \leq e^{-x}$ for all x .

Finally notice that $|\mathcal{D}| = \binom{n}{k} (2k^{k/2} + 1)^k \leq \left(\frac{ne}{k}\right)^k (ek^{k/2})^k \leq \left(\frac{e^2 nk^{k/2}}{k}\right)^k \leq e^{0.5k^2 \log n + 2k}$, where the inequalities are based on the fact that $k \geq 2$. By our assumption on the size of t and k , we therefore have $e^{-t^{1/2}} |\mathcal{D}| \leq (1/2)$. Therefore, taking a union bound over all $d \in \mathcal{D}$ of the previous displayed inequality gives $\Pr(\mathcal{E}) \geq 1/2$, concluding the proof of the lemma. \square

The next lemma takes care of the second step of the argument.

Lemma 4.5. *With probability at least $3/4$, the inequality $\sum_j x_j \leq \frac{n}{2} + 3\sqrt{n \log t}$ is valid for \mathbf{P} .*

Proof. Fix an $i \in [t]$. Since $\text{Var}(\mathbf{X}^i) = n/4$, we have from Bernstein's inequality

$$\begin{aligned} \Pr \left(\sum_j \mathbf{X}_j^i > \frac{n}{2} + 3\sqrt{n \log t} \right) &\leq \exp \left(-\min \left\{ 9 \log t, \frac{9\sqrt{n \log t}}{4} \right\} \right) \\ &\leq e^{-\frac{9 \log t}{4}} \leq \frac{1}{4t}, \end{aligned}$$

where the second inequality follows from the fact that $\log t \leq n$, and the last inequality uses the fact that $t \geq 4$. Taking a union bound over all $i \in [t]$ gives

$$\Pr \left(\bigvee_{i \in [t]} \left(\sum_j \mathbf{X}_j^i > \frac{n}{2} + 3\sqrt{n \log t} \right) \right) \leq \frac{1}{4},$$

Finally, notice that an inequality $dx \leq d_0$ is valid for \mathbf{P} iff it is valid for all \mathbf{X}^i . This concludes the proof. \square

Lemma 4.6. *Suppose that the polytope Q is α -tough for $\alpha \geq 1$ and that the inequality $\sum_i x_i \leq \frac{n}{2} + 3\sqrt{n \log t}$ is valid for Q . Then we have $d(Q, Q^k) \geq \sqrt{n} \left(\frac{\alpha}{2\sqrt{k}} - \frac{\alpha}{k^2} - \frac{3\sqrt{\log t}}{\sqrt{n}} \right)$.*

Proof. We first show that the point $\bar{q} = (\frac{1}{2} + \frac{\alpha}{2\sqrt{k}} - \frac{\alpha}{k^2})e$ belongs to Q^k . Let $dx \leq d_0$ be a facet for Q^k . Then we have

$$\begin{aligned} d\bar{q} &= \frac{\sum_i d_i}{2} + \alpha \left(\frac{1}{2\sqrt{k}} - \frac{1}{k^2} \right) \sum_i d_i \leq \frac{\sum_i d_i}{2} + \alpha \left(\frac{1}{2\sqrt{k}} - \frac{1}{k^2} \right) \|d\|_1 \\ &\leq \frac{\sum_i d_i}{2} + \alpha \left(\frac{1}{2\sqrt{k}} - \frac{1}{2k^2} \right) \|d\|_1 - \frac{\|d\|_\infty}{2k^2} \\ &\leq \frac{\sum_i d_i}{2} + \frac{\alpha}{2\sqrt{k}} \left(1 - \frac{1}{k^2} \right) \|d\|_1 - \frac{\|d\|_\infty}{2k^2}, \end{aligned}$$

where the first inequality uses the fact that $\frac{1}{2\sqrt{k}} - \frac{1}{k^2} \geq 0$ for $k \geq 2$ and the second inequality uses $\alpha \geq 1$ and $\|d\|_1 \geq \|d\|_\infty$. Since Q is α -tough it follows that \bar{q} satisfies $dx \leq d_0$; since this holds for all facets of Q^k , we have $\bar{q} \in Q^k$.

Now define the halfspace $H = \{x : \sum_i x_i \leq \frac{n}{2} + 3\sqrt{n \log t}\}$. By assumption $Q \subseteq H$, and hence $d(Q, Q^k) \geq d(H, Q^k)$. But it is easy to see that the point in H closest to \bar{q} is the point $\tilde{q} = (\frac{1}{2} + \frac{3\sqrt{\log t}}{\sqrt{n}})e$. This gives that $d(Q, Q^k) \geq d(H, Q^k) \geq d(\bar{q}, \tilde{q}) \geq \sqrt{n} \left(\frac{\alpha}{2\sqrt{k}} - \frac{\alpha}{k^2} - \frac{3\sqrt{\log t}}{\sqrt{n}} \right)$. This concludes the proof. \square

We now conclude the proof of Theorem 2.2.

Proof. of Theorem 2.2 Set $\bar{\alpha}^2 = \min \left\{ \frac{\log t}{12000 \log n}, \frac{k}{64} \right\}$. Taking union bound over Lemmas 4.4 and 4.5, with probability at least $1/4$, \mathbf{P} is $\bar{\alpha}$ -tough and the inequality $\sum_i x_i \leq \frac{n}{2} + 3\sqrt{n \log t}$ is valid for it. Then from Lemma 4.6 we get that with probability at least $1/4$, $d(\mathbf{P}, \mathbf{P}^k) \geq \sqrt{n} \left(\frac{\bar{\alpha}}{2\sqrt{k}} - \frac{\bar{\alpha}}{k^2} - \frac{3\sqrt{\log t}}{\sqrt{n}} \right)$, and the result follows by plugging in the value of $\bar{\alpha}$. \square

5 Hard Packing Integer Programs

In this section we prove Theorem 2.3. With overload in notation, we use $\binom{[n]}{k}$ to denote the set of vectors in $\{0, 1\}^n$ with exactly k 1's.

Let \mathbf{P} be a random polytope sampled from the distribution (n, m, M) -PIP and consider the corresponding random vectors \mathbf{A}^j 's. The idea of the proof is to show that with constant probability \mathbf{P} behaves like Example 2, by showing that the cut $\sum_i x_i \lesssim \frac{n}{2}$ is valid for it and that \mathbf{P} approximately contains 0/1 points with many 1's. Then we show that this 'approximate containment' implies that a point with a lot of mass (say, $\approx (1, 1, \dots, 1)$ for $k \leq n/2$) belongs to the k -sparse closure \mathbf{P}^k ; since such point is far from hyperplane $\sum_i x_i \lesssim \frac{n}{2}$, it is also far from \mathbf{P} and hence we get a lower bound on $d(\mathbf{P}, \mathbf{P}^k)$.

The first part of the argument is a straightforward application of Bernstein's inequality and union bound; its proof is presented in Section D of the appendix.

Lemma 5.1. *With probability at least $1 - \frac{1}{4}$ the cut $(1 - \frac{2\sqrt{\log 8n}}{\sqrt{m}}) \sum_i x_i \leq \frac{n}{2} + \frac{\sqrt{n \log 8}}{\sqrt{m}}$ is valid for \mathbf{P} .*

The other steps in the argument are more involved.

5.1 Approximate Containment of Points with Many 1's

First we control the right-hand side of the constraints $\mathbf{A}^j x \leq \frac{\sum_i \mathbf{A}_i^j}{2}$ that define \mathbf{P} , by showing that they are roughly $\frac{nM}{2}$; this is again a straightforward application of Bernstein's inequality and is also deferred to Section D of the appendix.

Lemma 5.2. *With probability at least $1 - \frac{1}{8}$ we have $|\sum_{i=1}^n \mathbf{A}_i^j - \frac{nM}{2}| \leq M\sqrt{n \log 8m}$ for all $j \in [m]$.*

Recall that we defined $c = \frac{k}{8}$. Now we show that with constant probability, *all* points $\bar{x} \in \{0, 1\}^n$ with cn 1's satisfy $\mathbf{A}^j \bar{x} \lesssim \frac{nM}{2}$ for all $j \in [m]$, and hence they approximately belong to \mathbf{P} . The argument is cleaner if the random variables \mathbf{A}_i^j were uniformly distributed in the *continuous* interval $[0, M]$, instead of on the discrete set $\{0, \dots, M\}$; this is because in the former we can leverage the knowledge of the *order statistics* of continuous uniform variables. Our next lemma then essentially handles this continuous case.

Lemma 5.3. *Let $\mathbf{U} \in \mathbb{R}^n$ be a random variable where each coordinate U_i is independently drawn uniformly from $[0, 1]$. Then with probability at least $1 - 1/8m$ we have $\mathbf{U} \bar{x} \leq \frac{c(2n - cn + 1)}{2} + \sqrt{10cnm}$ for all vectors $\bar{x} \in \binom{[n]}{cn}$.*

Proof. Let $U_{(i)}$ be the i th order statistics of U_1, U_2, \dots, U_n (i.e. in each scenario $U_{(i)}$ equals the i th smallest value among U_1, U_2, \dots, U_n in that scenario). Notice that $\max_{\bar{x} \in \binom{[n]}{cn}} \mathbf{U} \bar{x} = U_{(n)} + \dots + U_{(n - cn + 1)}$, and hence it is equivalent to show that

$$\Pr \left(U_{(n)} + \dots + U_{(n - cn + 1)} > \frac{c(2n - cn + 1)}{2} + \sqrt{10cnm} \right) \leq \frac{1}{8m}.$$

We use $\mathbf{Z} \triangleq U_{(n)} + \dots + U_{(n - cn + 1)}$ to simplify the notation.

It is known that $\mathbb{E}[\mathbf{U}_{(i)}] = \frac{i}{n+1}$ and $\text{Cov}(\mathbf{U}_{(i)}, \mathbf{U}_{(j)}) = \frac{i(n+1-j)}{(n+1)^2(n+2)} \leq \frac{1}{n}$ [5]. Also, since $\mathbf{U}_{(i)}$ lies in $[0, 1]$, we have $\text{Var}(\mathbf{U}_{(i)}) \leq 1/4$. Using this information, we get $\mathbb{E}[\mathbf{Z}] = \frac{(2n-cn+1)cn}{2(n+1)} \leq \frac{c(2n-cn+1)}{2}$ and

$$\text{Var}(\mathbf{Z}) \leq \frac{cn}{4} + \frac{(cn)^2}{n} \leq \frac{5cn}{4},$$

where the last inequality follows from the fact $c \leq 1$. Then applying Chebychev's inequality [8], we get

$$\Pr\left(\mathbf{Z} \geq \frac{c(2n-cn+1)}{2} + \sqrt{10cnm}\right) \leq \frac{\text{Var}(\mathbf{Z})}{10cnm} \leq \frac{1}{8m}.$$

This concludes the proof. \square

Now we translate this proof from the continuous to the discrete setting.

Lemma 5.4. *With probability at least $1 - \frac{1}{8}$ we have*

$$\mathbf{A}^j \bar{x} \leq \frac{(M+1)c(2n-cn+1)}{2} + (M+1)\sqrt{10cnm}, \quad \forall j \in [m], \forall \bar{x} \in \binom{[n]}{cn}.$$

Proof. For each $j \in [m]$, let $\mathbf{U}_1^j, \mathbf{U}_2^j, \dots, \mathbf{U}_n^j$ be independent and uniformly distributed in $[0, 1]$. Define $\mathbf{Y}_i^j \triangleq \lfloor (M+1)\mathbf{U}_i^j \rfloor$. Notice that the random variables $(\mathbf{Y}_i^j)_{i,j}$ have the same distribution as $(\mathbf{A}_i^j)_{i,j}$. So it suffices to prove the lemma for the variables \mathbf{Y}_i^j 's.

Fix $j \in [m]$. For any $\bar{x} \in \{0, 1\}^n$ we have $\mathbf{Y}^j \bar{x} \leq (M+1)\mathbf{U} \bar{x}$. Therefore, from Lemma 5.3 we get

$$\Pr\left(\bigvee_{\bar{x} \in \binom{[n]}{cn}} \left(\mathbf{Y}^j \bar{x} > \frac{(M+1)c(2n-cn+1)}{2} + (M+1)\sqrt{10cnm}\right)\right) \leq \frac{1}{8m}.$$

Taking a union bound of this last expression over all $j \in [m]$ concludes the proof of the lemma. \square

5.2 From Approximate to Actual Containment

From the previous section we get with constant probability, points $\bar{x} \in \{0, 1\}^n$ with cn 1's approximately belong to \mathbf{P} ; thus, scaling them by a small factor, shows that these points belong to the *LP relaxation* of \mathbf{P} . Our goal is to strengthen this result by showing that a small (although slightly larger) scaling of these point actually brings them to the integer hull \mathbf{P} itself.

The next lemma shows that this is in fact possible.

Lemma 5.5. *Consider a 0/1 polytope $Q = \text{conv}(\{x \in \{0, 1\}^n : a^j x \leq b_j, j = 1, 2, \dots, m\})$ where $n \geq 50$, $m \leq n$, $a_i^j \in [0, M]$ for all i, j , and $b_j \geq \frac{nM}{12}$ for all j . Consider $1 < \alpha \leq 2\sqrt{n}$ and let $\bar{x} \in \{0, 1\}^n$ be such that for all j , $a^j \bar{x} \leq \alpha b_j$. Then the point $\frac{1}{\alpha}(1 - \epsilon)^2 \bar{x}$ belongs to Q as long as $\frac{12\sqrt{\log 4n^2m}}{\sqrt{n}} \leq \epsilon \leq \frac{1}{2}$.*

For the remainder of the section we prove this lemma. The idea is that we can select a subset of $\approx 1 - 1/\alpha$ coordinates and change \bar{x} to 0 in these coordinates to obtain a feasible solution in Q ; repeating this for many sets of coordinates and taking an average of the feasible points obtained will give the result.

To make this precise, let $p = \frac{1}{\alpha}(1 - \epsilon)$. For $w \in [n^2]$ define the independent random variables $\mathbf{X}_1^w, \mathbf{X}_2^w, \dots, \mathbf{X}_n^w$ taking values in $\{0, 1\}$ such that $\mathbb{E}[\mathbf{X}_i^w] = p\bar{x}_i$ (i.e. if $\bar{x}_i = 1$, then keep it at 1 with probability p , otherwise flip it to 0; if $\bar{x}_i = 0$, then keep it at 0).

Claim 5.6. *With probability at least $3/4$ all points \mathbf{X}^w belong to Q .*

Proof. Notice $\mathbb{E}[a^j \mathbf{X}^w] \leq (1 - \epsilon)b_j$. Also, from our upper bound on a^j , we have $\text{Var}(a^j \mathbf{X}^w) \leq \frac{M^2 n}{4}$. Employing Bernstein's inequality,

$$\Pr(a^j \mathbf{X}^w > b_j) \leq \exp \left(- \min \left\{ \frac{\epsilon^2 b_j^2}{M^2 n}, \frac{3\epsilon b_j}{4M} \right\} \right) \leq \frac{1}{4n^2 m},$$

where the second inequality uses the assumed lower bounds on b_j and ϵ , and the fact that $\frac{4 \cdot 12 \log 4n^2 m}{3n} \leq \frac{12\sqrt{\log 4n^2 m}}{\sqrt{n}}$ due to our bounds on n and m . The claim follows by taking a union bound over all j and w . \square

Let $\mathbf{Z} = \frac{1}{n^2} \sum_w \mathbf{X}^w$ be the random point that is the average of the \mathbf{X}^w 's.

Claim 5.7. *With probability at least $3/4$, $\mathbf{Z}_i \geq \frac{1}{\alpha}(1 - \epsilon)^2 \bar{x}_i$ for all i .*

Proof. Since $\bar{x} \in \{0, 1\}^n$, it suffices to consider indices i such that $\bar{x}_i = 1$. Fix such an i . We have $\mathbb{E}[n^2 \mathbf{Z}_i] = pn^2$ and $\text{Var}(n^2 \mathbf{Z}_i) \leq \frac{n^2}{4}$. Then from Bernstein's inequality

$$\begin{aligned} \Pr \left(\mathbf{Z}_i < \frac{1}{\alpha}(1 - \epsilon)^2 \bar{x}_i \right) &= \Pr(n^2 \mathbf{Z}_i < \mathbb{E}[n^2 \mathbf{Z}_i](1 - \epsilon)) \\ &\leq \exp \left(- \min \left\{ n^2(\epsilon p)^2, \frac{3n^2 \epsilon p}{4} \right\} \right) \leq \frac{1}{4n}, \end{aligned}$$

where the last inequality uses the lower bound on ϵ , the fact that $n \geq 50$, and the fact that $p \geq 1/2\alpha \geq 1/4\sqrt{n}$. The claim follows from taking a union bound over all i such that $\bar{x}_i = 1$. \square

Taking a union bound over the above two claims we get that there is a realization $\tilde{x}^1, \tilde{x}^2, \dots, \tilde{x}^{n^2}$ of the random vectors $\mathbf{X}^1, \mathbf{X}^2, \dots, \mathbf{X}^{n^2}$ such that (let $\tilde{z} = \frac{1}{n^2} \sum_w \tilde{x}^w$): (i) All \tilde{x}^w belong to Q , and hence so does their convex combination \tilde{z} ; (ii) $\tilde{z} \geq \frac{1}{\alpha}(1 - \epsilon)^2 \bar{x}$. Since Q is of packing-type, it follows that the point $\frac{1}{\alpha}(1 - \epsilon)^2 \bar{x}$ belongs to Q , concluding the proof of Lemma 5.5.

5.3 Proof of Theorem 2.3

Now we put together the results from the previous sections to conclude the proof of Theorem 2.3. Let \mathcal{E} be the event that Lemmas 5.2, 5.1 and 5.4 hold; notice that $\Pr(\mathcal{E}) \geq 1/2$. For the rest of the proof we fix a \mathbf{P} (and the associated \mathbf{A}^j 's) where \mathcal{E} holds and prove a lower bound on $d(\mathbf{P}, \mathbf{P}^k)$.

Consider a set $I \in \binom{[n]}{cn}$ and let \bar{x} be the incidence vector of I (i.e. $\bar{x}_i = 1$ if $i \in I$ and $\bar{x}_i = 0$ if $i \notin I$). Since the bounds from Lemmas 5.2 and 5.4 hold for our \mathbf{P} , straightforward calculations show that $\mathbf{A}^j \bar{x} \leq \alpha \frac{1}{2} \sum_i \mathbf{A}_i^j$ for all $j \in [m]$. Therefore, from Lemma 5.5 we have that the point

$\frac{1}{\max\{\alpha, 1\}}(1 - \epsilon)^2 \tilde{x}$ belongs to \mathbf{P} . This means that the point $\tilde{x} = \frac{1}{\max\{\alpha, 1\}}(1 - \epsilon)^2 e$ belongs to $\mathbf{P} + \mathbb{R}^{\bar{I}}$ (see Section 1.1). Since this holds for every $I \in \binom{[n]}{cn}$, we have $\tilde{x} \in \mathbf{P}^k$.

Let $\tilde{\mathbf{y}}$ be the point in \mathbf{P} closest to \tilde{x} . Let $a = (1 - \frac{2\sqrt{\log 8n}}{\sqrt{m}})$ and $b = \frac{n}{2} + \sqrt{n \log 8m}$, so that the cut in Lemma 5.1 is given by $ae\mathbf{x} \leq b$. From Cauchy-Schwarz we have that $d(\tilde{x}, \tilde{\mathbf{y}}) \geq \frac{ae\tilde{x} - ae\tilde{\mathbf{y}}}{\|ae\|} = \frac{e\tilde{x}}{\sqrt{n}} - \frac{ae\tilde{\mathbf{y}}}{a\sqrt{n}}$.

By definition of \tilde{x} we have $e\tilde{x} = \frac{1}{\max\{\alpha, 1\}}(1 - \epsilon)^2 n$. From the fact the cut $ae\mathbf{x} \leq b$ is valid for \mathbf{P} and $\tilde{\mathbf{y}} \in \mathbf{P}$, we have $ae\tilde{\mathbf{y}} \leq b$. Simple calculations show that $\frac{b}{a\sqrt{n}} \leq \frac{n}{2}(1 + \epsilon')$. Plugging these values in we get that $d(\mathbf{P}, \mathbf{P}^k) = d(\tilde{x}, \tilde{\mathbf{y}}) \geq \frac{\sqrt{n}}{2} \left(\frac{2(1 - \epsilon)^2}{\max\{\alpha, 1\}} - (1 + \epsilon') \right)$. Theorem 2.3 follows from the definition of α, ϵ and ϵ' .

6 Sparse Cutting-Planes and Extended Formulations

In this section we analyze the relationship between sparse cuts and extended formulations, proving Proposition 2.4 and Proposition 2.5.

6.1 Proof of Proposition 2.4

For any set $S \subseteq \mathbb{R}^{n'}$ and $I \subseteq [n']$, define $\tau_I(S) = S + \mathbb{R}^{\bar{I}}$ (recall that $\mathbb{R}^{\bar{I}} = \{x \in \mathbb{R}^{n'} : x_i = 0 \text{ for } i \in I\}$).

Consider $P \subseteq \mathbb{R}^n$ and $Q \subseteq \mathbb{R}^n \times \mathbb{R}^m$ such that $P = \text{proj}_x(Q)$. Given a subset $I \subseteq [n + m]$ we use I_x to denote the indices of I in $[n]$ (i.e. $I_x = I \cap [n]$). We start with the following technical lemma.

Lemma 6.1. *For every $I \subseteq [n + m]$ we have $\tau_{I_x}(\text{proj}_x(Q)) = \text{proj}_x(\tau_I(Q))$.*

Proof. (\subseteq) Take $u_x \in \tau_{I_x}(\text{proj}_x(Q))$; this means that there is $v \in Q$ such that $u_x = \text{proj}_x(v) + d_x$ for some vector $d_x \in \mathbb{R}^n$ with support in I_x . Define $d = (d_x, 0) \in \mathbb{R}^n \times \mathbb{R}^m$, with support in $I_x \subseteq I$. Then $v + d$ belongs to $\tau_I(Q)$ and

$$u_x = \text{proj}_x(v) + d_x = \text{proj}_x(v + d) \in \text{proj}_x(\tau_I(Q)),$$

concluding this part of the proof.

(\supseteq) Take $u_x \in \text{proj}_x(\tau_I(Q))$. Let $u \in \tau_I(Q)$ be such that $\text{proj}_x(u) = u_x$. By definition, there is $d \in \mathbb{R}^n \times \mathbb{R}^m$ with support in I such that $u + d$ belongs to Q . Then $\text{proj}_x(u + d) = u_x + \text{proj}_x(d)$ belongs to $\text{proj}_x(Q)$; since $\text{proj}_x(d)$ is supported in I_x , we have that u_x belongs to $\tau_{I_x}(\text{proj}_x(Q))$, thus concluding the proof of the lemma. \square

The proof of Proposition 2.4 then follows directly from the above lemma:

$$\begin{aligned} (\text{proj}_x(Q))^k &= \bigcap_{J \subseteq [n]} \tau_J(\text{proj}_x(Q)) = \bigcap_{I \subseteq [n+m]} \tau_{I_x}(\text{proj}_x(Q)) \\ &\stackrel{\text{Lemma 6.1}}{=} \bigcap_{I \subseteq [n+m]} \text{proj}_x(\tau_I(Q)) \supseteq \text{proj}_x \left(\bigcap_{I \subseteq [n+m]} \tau_I(Q) \right) = \text{proj}_x(Q^k). \end{aligned}$$

6.2 Proof of Proposition 2.5

We construct the polytope $Q \subseteq \mathbb{R}^n \times \mathbb{R}^{2n-1}$ as follows. Let T be the complete ordered binary tree of height $\ell + 1$. We let r denote the root node of T . We use $\text{int}(T)$ to denote the set of internal nodes of T , and for an internal node $v \in \text{int}(T)$ we use $\text{left}(v)$ to denote its left child and $\text{right}(v)$ to denote its right child. Let $i(\cdot)$ be a bijection between the leaf nodes of T and the elements of $[n]$. We then define the set Q as the solutions (x, y) to the following:

$$\begin{aligned} y_r &\leq 1 \\ y_v &= y_{\text{left}(v)} + y_{\text{right}(v)}, \quad \forall v \in \text{int}(T) \\ y_v &= \frac{2}{n} x_{i(v)}, \quad \forall v \in T \setminus \text{int}(T) \\ y_v &\geq 0, \quad \forall v \in T \\ x_i &\in [0, 1], \quad \forall i \in [n]. \end{aligned} \tag{2}$$

Define $P = \{x \in [0, 1]^n : \sum_{i \in [n]} x_i \leq n/2\}$.

Claim 6.2. Q is an extended formulation of P , namely $\text{proj}_x(Q) = P$.

Proof. (\subseteq) Take $(\bar{x}, \bar{y}) \in Q$. Let T_j denote the set of nodes of T at level j . It is easy to see (for instance, by reverse induction on j) that $\sum_{v \in T_j} \bar{y}_v = \frac{2}{n} \sum_{i \in [n]} \bar{x}_i$ for all j . In particular, $\bar{y}_r = \frac{2}{n} \sum_{i \in [n]} \bar{x}_i$. Since $\bar{y}_r \leq 1$, we have that $\bar{x} \in P$.

(\supseteq) Take $\bar{x} \in P$. Define \bar{y} inductively by setting $\bar{y}_{i(v)} = \frac{2}{n} \bar{x}_{i(v)}$ for all leaves v and $\bar{y}_v = \bar{y}_{\text{left}(v)} + \bar{y}_{\text{right}(v)}$ for all internal nodes v . As in the previous paragraph, it is easy to see that $\bar{y}_r = \sum_{i \in [n]} \bar{x}_i \leq n/2$. Therefore, (\bar{x}, \bar{y}) belongs to Q . \square

Claim 6.3. $d(P, P^k) = \sqrt{n/2}$ for all $k \leq n/2$.

Proof. For every subset $I \subseteq [n]$ of size $n/2$, the incidence vector of I belongs to P this implies that, when $k \leq n/2$, the all ones vector e belongs to P^k . It is easy to see that the closest vector in P to e is the vector $\frac{1}{2}e$; since the distance between e and $\frac{1}{2}e$ is $\sqrt{n/2}$, the claim follows. \square

Claim 6.4. $Q^3 = Q$.

Proof. Follows directly from the fact that all the equations and inequalities defining Q in (2) have support of size at most 3. \square

The proof of Proposition 2.5 follows directly from the three claims above.

References

- [1] T. Achterberg. Personal communication.
- [2] K. Andersen and R. Weismantel. Zero-coefficient cuts. In *IPCO*, 2010.
- [3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [4] A. DasGupta. *Probability for Statistics and Machine Learning*. Springer-Verlag, 2011.

- [5] H. David and H. Nagaraja. *Order Statistics*. Wiley, 2003.
- [6] Z. Gu. Personal communication.
- [7] K. Kaparis and A. N. Letchford. Separation algorithms for 0-1 knapsack polytopes. *Mathematical Programming*, 124(1-2):69–91, 2010.
- [8] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer-Verlag, 2011.
- [9] J. Matousek and J. Vondrak. *The Probabilistic Method*. 2008. Manuscript.
- [10] A. Narisetty. Personal communication.
- [11] G. M. Ziegler. Lectures on 0/1-polytopes. In *Polytopes combinatorics and computation*, pages 1–41. Springer, 2000.

A Concentration Inequalities

We state Bernstein’s inequality in a slightly weaker but more convenient form.

Theorem A.1 (Bernstein’s Inequality [[8], Appendix A.2]). *Let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ be independent random variables such that $\|\mathbf{X}_i - \mathbb{E}[\mathbf{X}_i]\| \leq M$ for all $i \in [n]$. Let $\mathbf{X} = \sum_{i=1}^n \mathbf{X}_i$ and define $\sigma^2 = \text{Var}(\mathbf{X})$. Then for all $t > 0$ we have*

$$\Pr(\|\mathbf{X} - \mathbb{E}[\mathbf{X}]\| > t) \leq \exp \left(- \min \left\{ \frac{t^2}{4\sigma^2}, \frac{3t}{4M} \right\} \right).$$

B Empirically Generating Lower Bound on $d(P, P^k)$

We estimate a lower bound on $d(P, P^k)$ using the following procedure. The input to the procedure is the set of points $\{p^1, \dots, p^t\} \in [0, 1]^n$ which are vertices of P . For every $I \in \binom{[n]}{k}$, we use PORTA to obtain an inequality description of $P + \mathbb{R}^{\bar{I}}$. Putting all these inequalities together we obtain an inequality description of P^k . Unfortunately due to the large number of inequalities, we are unable to find the vertices of P^k using PORTA. Therefore, we obtain a lower bound on $d(P, P^k)$ via a *shooting experiment*.

First observe that given $u \in \mathbb{R}^n \setminus \{0\}$ we obtain a lower bound on $d(P, P^k)$ as:

$$\frac{1}{\|u\|} \left(\max\{u^T x : x \in P^k\} - \max\{u^T x : x \in P\} \right).$$

Moreover it can be verified that there exists a direction which achieves the correct value of $d(P, P^k)$. We generated 20,000 random directions u by picking them uniformly in the set $[-1, 1]^n$. Also we found that for instances where $p^j \in \{x \in \{0, 1\}^n : \sum_{i=1}^n x_i = \frac{n}{2}\}$, the directions $(\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}})$ and $-(\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}})$ yield good lower bounds. The Figure in Section 1.3(c) plots the best lower bound among the 20,002 lower bounds found as above.

C Anticoncentration of Linear Combination of Bernoulli's

It is convenient to restate Lemma 4.2 in terms of Rademacher random variables (i.e. that takes values $-1/1$ with equal probability).

Lemma C.1 (Lemma 4.2, restated). *Let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ be independent Rademacher random variables. Then for every $a \in [-1, 1]^n$,*

$$\Pr\left(a\mathbf{X} \geq \frac{\alpha}{\sqrt{n}}\left(1 - \frac{1}{n^2}\right)\|a\|_1 - \frac{1}{n^2}\right) \geq \left(e^{-50\alpha^2} - e^{-100\alpha^2}\right)^{60\log n}, \quad \alpha \in \left[0, \frac{\sqrt{n}}{8}\right].$$

We start with the case where the vector a has all of its coordinates being similar.

Lemma C.2. *Let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ be independent Rademacher random variables. For every $\epsilon \geq 1/20$ and $a \in [1 - \epsilon, 1]^n$,*

$$\Pr\left(a\mathbf{X} \geq \frac{\alpha}{\sqrt{n}}\|a\|_1\right) \geq e^{-50\alpha^2} - e^{-\frac{\alpha^2}{4\epsilon^2}}, \quad \alpha \in \left[0, \frac{\sqrt{n}}{8}\right].$$

Proof. Since $a\mathbf{X} = \sum_i \mathbf{X}_i - \sum_i (1 - a_i)\mathbf{X}_i$, having $\sum_i \mathbf{X}_i \geq 2t$ and $\sum_i (1 - a_i)\mathbf{X}_i \leq t$ implies that $a\mathbf{X} \geq t$. Therefore,

$$\begin{aligned} \Pr(a\mathbf{X} \geq t) &\geq \Pr\left(\left(\sum_i \mathbf{X}_i \geq 2t\right) \vee \left(\sum_i (1 - a_i)\mathbf{X}_i \leq t\right)\right) \\ &\geq \Pr\left(\sum_i \mathbf{X}_i \geq 2t\right) - \Pr\left(\sum_i (1 - a_i)\mathbf{X}_i \leq t\right), \end{aligned} \quad (3)$$

where the second inequality comes from union bound. For $t \in [0, n/8]$, the first term in the right-hand side can be lower bounded by $e^{-\frac{50t^2}{n}}$ (see for instance Section 7.3 of [9]). The second term in the right-hand side can be bounded using Bernstein's inequality: given that $\text{Var}(\sum_i (1 - a_i)\mathbf{X}_i) = \sum_i (1 - a_i)^2 \leq n\epsilon^2$, we get that for all $t \in [0, n/8]$

$$\Pr\left(\sum_i (1 - a_i)\mathbf{X}_i \leq t\right) \leq \exp\left(-\min\left\{\frac{t^2}{4n\epsilon^2}, \frac{3t}{4\epsilon}\right\}\right) = e^{-\frac{t^2}{4n\epsilon^2}}.$$

The lemma then follows by plugging these bounds on (3) and using $t = \alpha\sqrt{n} \geq \frac{\alpha}{\sqrt{n}}\|a\|_1$. \square

Proof of Lemma C.1. Without loss of generality assume $a > 0$, since flipping the sign of negative coordinates of a does not change the distribution of $a\mathbf{Z}$ neither the term $\frac{\alpha}{\sqrt{n}}\left(1 - \frac{2}{n^2}\right)\|a\|_1$. Also assume without loss of generality that $\|a\|_\infty = 1$. The idea of the proof is to bucket the coordinates such that in each bucket the values of a is within a factor of $(1 \pm \epsilon)$ of each other, and then apply Lemma C.2 in each bucket.

The first step is to trim the coefficients of a that are very small. Define the trimmed version b of a by setting $b_i = a_i$ for all i where $a_i \geq 1/n^3$ and $b_i = 0$ for all other i . We first show that

$$\Pr\left(b\mathbf{Z} \geq \frac{\alpha}{\sqrt{n}}\|b\|_1\right) \geq \left(e^{-50\alpha^2} - e^{-100\alpha^2}\right)^{60\log n}, \quad (4)$$

and then we argue that the error introduced by considering b instead of a is small.

For $j \in \{0, 1, \dots, \frac{3 \log n}{\epsilon}\}$, define the j th bucket as $I_j = \{i : b_i \in ((1 - \epsilon)^{j+1}, (1 - \epsilon)^j]\}$. Since $(1 - \epsilon)^{\frac{3 \log n}{\epsilon}} \leq e^{-3 \log n} = 1/n^3$, we have that every index i with $b_i > 0$ lies within some bucket.

Now fix some bucket j . Let $\epsilon = 1/20$ and $\gamma = \frac{\alpha}{\sqrt{n}}$. Let E_j be the event that $\sum_{i \in I_j} b_i \mathbf{Z}_i \geq \gamma \sum_{i \in I_j} b_i$. Employing Lemma C.2 over the vector $(1 - \epsilon)^j b|_{I_j}$, gives

$$\Pr \left(\sum_{i \in I_j} b_i \mathbf{Z}_i \geq \gamma \sum_{i \in I_j} b_i \right) \geq e^{-50\gamma^2 |I_j|} - e^{-\frac{\gamma^2 |I_j|}{4\epsilon^2}} \geq e^{-50\gamma^2 n} - e^{-\frac{\gamma^2 n}{4\epsilon^2}}, \quad \gamma \in \left[0, \frac{1}{8}\right].$$

But now notice that if in a scenario we have E_j holding for all j , then in this scenario we have $b\mathbf{Z} \geq \gamma \|b\|_1$. Using the fact that the E_j 's are independent (due to the independence of the coordinates of \mathbf{Z}), we have

$$\Pr(b\mathbf{Z} \geq \gamma \|b\|_1) \geq \Pr \left(\bigvee_j E_j \right) \geq \left(e^{-50\gamma^2 n} - e^{-\frac{\gamma^2 n}{4\epsilon^2}} \right)^{\frac{3 \log n}{\epsilon}}, \quad \gamma \in \left[0, \frac{1}{8}\right].$$

Now we claim that whenever $b\mathbf{Z} \geq \gamma \|b\|_1$, then we have $a\mathbf{Z} \geq \frac{\alpha}{\sqrt{n}} \left(1 - \frac{2}{n^2}\right) \|a\|_1$. First notice that $\|b\|_1 \geq \|a\|_1 - 1/n^2 \geq \|a\|_1(1 - 1/n^2)$, since $\|a\|_1 \geq \|a\|_\infty = 1$. Moreover, with probability 1 we have $a\mathbf{Z} \geq b\mathbf{Z} - 1/n^2$. Therefore, whenever $b\mathbf{Z} \geq \gamma \|b\|_1$:

$$a\mathbf{Z} \geq b\mathbf{Z} - \frac{1}{n^2} \geq \gamma \|b\|_1 - \frac{1}{n^2} \geq \gamma \left(1 - \frac{1}{n^2}\right) \|a\|_1 - \frac{1}{n^2} = \frac{\alpha}{\sqrt{n}} \left(1 - \frac{1}{n^2}\right) \|a\|_1 - \frac{1}{n^2}.$$

This concludes the proof of the lemma. \square

D Hard Packing Integer Programs

D.1 Proof of Lemma 5.1

Fix $i \in [n]$. We have $\mathbb{E}[\sum_j \mathbf{A}_i^j] = \frac{mM}{2}$ and $\text{Var}(\sum_j \mathbf{A}_i^j) \leq \frac{mM^2}{4}$. Employing Bernstein's inequality we get

$$\Pr \left(\sum_j \mathbf{A}_i^j < \frac{mM}{2} - \sqrt{m \log 8nM} \right) \leq \exp \left(- \min \left\{ \log 8n, \frac{3\sqrt{m \log 8n}}{4} \right\} \right) \leq \frac{1}{8n},$$

where the last inequality uses the assumption that $m \geq 8 \log 8n$. Similarly, we get that

$$\Pr \left(\sum_{i,j} \mathbf{A}_i^j > \frac{nmM}{2} + \sqrt{nm \log 8nM} \right) \leq \exp \left(- \min \left\{ \log 8n, \frac{3\sqrt{nm \log 8n}}{4} \right\} \right) \leq \frac{1}{8n}.$$

Taking a union bound over the first displayed inequality over all $i \in [n]$ and also over the last inequality, with probability at least $1 - 1/4$ the valid cut $\sum_i \left(\frac{2}{mM} \sum_j \mathbf{A}_i^j \right) x_i \leq \frac{1}{mM} \sum_{i,j} \mathbf{A}_i^j$ (obtained by aggregating all inequalities in the formulation) has all coefficients on the left-hand side being at least $(1 - \frac{2\sqrt{\log 8n}}{\sqrt{m}})$ and the right-hand side at most $\frac{n}{2} + \frac{\sqrt{n \log 8}}{\sqrt{m}}$. This concludes the proof.

D.2 Proof of Lemma 5.2

Fix $j \in [m]$. We have $\mathbb{E}[\sum_{i=1}^n \mathbf{A}_i^j] = \frac{nM}{2}$ and $\text{Var}(\sum_{i=1}^n \mathbf{A}_i^j) \leq nM^2/4$ and hence by Bernstein's inequality we get

$$\Pr\left(\sum_{i=1}^n \mathbf{A}_i^j > \frac{nM}{2} + M\sqrt{n \log 8m}\right) \leq \exp\left(-\min\left\{\log 8m, \frac{3\sqrt{n \log 8m}}{4}\right\}\right) \leq \frac{1}{8m},$$

where the last inequality uses the assumption that $m \leq n$. The lemma then follows by taking a union bound over all $j \in [m]$.